Available online at www.sciencedirect.com

SciVerse ScienceDirect

Physics Procedia 33 (2012) 1518 – 1525

Physics Procedia

2012 International Conference on Medical Physics and Biomedical Engineering

Study of Selective Ensemble Learning Methods Based on Support Vector Machine

Kai Li, Zhibin Liu ,Yanxia Han

School of Mathematics and Computer Hebei University
Baoding, Hebei Province,071002, China
likai_njtu@163.com,chenglong_lzb@163.com

Abstract

Diversity among base classifiers is an important factor for improving in ensemble learning performance. In this paper, we choose support vector machine as base classifier and study four methods of selective ensemble learning which include hill-climbing, ensemble forward sequential selection, ensemble backward sequential selection and clustering selection. To measure the diversity among base classifiers in ensemble learning, the entropy E is used. The experimental results show that different diversity measure impacts on ensemble performance in some extent and first three selective strategies have similar generalization performance. Meanwhile, when using clustering selective strategy, selecting different number of clusters in this experiment also does not impact on the ensemble performance except some dataset.

Keywords: Diversity; Selective Ensemble; Generalization Error; Support Vector Machine

1. Introduction

Diversity has been recognized as a very important characteristic for improving generalization performance of ensemble learning. So researchers present some diversity measures and ensemble learning methods which use different strategy to raise diversity among components. In generating ensemble member, selective ensemble learning is a common method. Using different selective strategies will obtain different ensemble learning methods. For example, aimed at neural network, Giacinto applied clustering technology to select ensemble member [1]. Zhou et al utilized genetic algorithm to select ensemble members and obtained better generalization performance [2]. After that, Li et al. also applied clustering techniques and genetic algorithms to select ensemble models [3]. Moreover, we study the selective ensemble learning based on neural network and decision tree [4]. In the aspect of diversity,

 doi:10.1016/j.phpro.2012.05.247

Kuncheva et al researched the measures of diversity in classifier ensembles and their relationship with the ensemble accuracy [5]. Now, people are still researching the measures of diversity and different ensemble learning algorithms [6-13]. Aimed these above, this paper researches the selective ensemble methods based on support vector machine algorithms.

This paper is organized as follows. Support vector machine and diversity measure used in this paper are summarized in section 2, and selective ensemble methods are introduced in section 3. Section 4 gives the results and analysis of experiment. The conclusions are given in section 5.

2. Support vector machine and diversity measure

In this section, we briefly review the support vector machine in binary classification problems.Given a dataset of labeled training points (x_1, y_1), (x_2, y_2),…, (x_l, y_l), where $(x_i, y_i) \subseteq R^N \times \{+1, -1\}$, i=1, 2…l. When they are linearly separable training dataset, there exist some hyperplane which correctly separate the positive and negative examples. The point x which lies on the hyperplane satisfies $<w,x>$+b=0, where w is normal to the hyperplane. It is seen that if the training set is linearly separable, the support vector algorithm finds the optimal separating hyperplane with the maximal margin; If the training set is linearly non-separable or approximately separable data, it need introduce the trade-off parameter; If the training data is not linearly separable, the SVM leaning algorithm mapped the input data using a nonlinearly mapping function $x \rightarrow \varphi(x)$ to a high-dimension feature space z, and the data in z is indeed linearly or approximately separable. In two-class classification, all training data satisfy the following decision function

$$f(x_i) = sign(< w, x > +b) = \begin{cases} +1, & if \ \mathrm{y_i}=+1 \\ -1, & if \ \mathrm{y_i}=-1 \end{cases}. \tag{1}$$

In the linearly separable training set, all training points satisfy the following inequalities

$$\begin{cases} < w, x_i > +b \geq +1, & \text{if } \mathrm{y_i}=+1 \\ < w, x_i > +b \leq -1, & \text{if } \mathrm{y_i}=-1 \end{cases}. \tag{2}$$

In fact, it can be written as $\mathrm{y_i}(< w, x_i > +b) \geq 1$, ($i$=1,2,…,$l$) above inequalities. Finding the hyperplane is equivalent to obtain the maximum margin by minimizing $\| w \|^2$, subject to constraints (2). The primal optimal problem is given as

$$\begin{aligned} &\min_{w,b} \ \frac{1}{2} \| w \|^2 \\ &s.t. \ \ \mathrm{y_i}(<\mathrm{w},\mathrm{x_i}>+\mathrm{b}) \geq 1, \text{for i=1,2,...,l} \end{aligned} \tag{3}$$

As the process of solving (3) is very different, so we introduce Lagrange multiplier to transform the primal problem into its dual problem that solves the following quadratic programming (QP) problem.

$$\begin{aligned} &\min_{\alpha} \ \ \frac{1}{2} \sum_{i=1}^{l} \sum_{j=1}^{l} \alpha_i \alpha_j y_i y_j (x_i \cdot x_j) - \sum_{i=1}^{l} \alpha_i \\ &s.t. \ \ \sum_{i=1}^{l} \alpha_i y_i = 0, \ 0 \leq \alpha_i, i = 1, 2, \ldots, l. \end{aligned} \tag{4}$$

In linear classifier, the solution in feature space using a linearly mapping function $X \rightarrow \varphi(X)$ only replaces the dot products $x \cdot x_j$ by inner product of vectors $\varphi(x) \cdot \varphi(x_j)$. The mapping function satisfies $< \varphi(x), \varphi(x_j) >= k(x, x_i)$, called kernel function, in the training algorithm and we would never need to explicitly even know what φ is. An decision function SVM is obtained by computing dot products of a given test point x with w, or more specifically by computing the sign of

$$\begin{aligned} f(x) &= \sum_{i=1}^{Ns} \alpha^*_i y_i (s_i \cdot x) + b \\ &= \sum_{i=1}^{Ns} \alpha^*_i y_i \varphi(s_i) \cdot \varphi(x) + b \\ &= \sum_{i=1}^{Ns} \alpha^*_i y_i K(s_i, x) + b \end{aligned}, \tag{5}$$

where the coefficients $\{\alpha_i\}$ are positive and subtracted from the objective function, the s_i are support vectors, and Ns is the number of support vectors.

Next we breifly introduce Non-pairwise diversity measure used in the experiment, which is the entropy measure E [6], which is defined in the following:

$$E = \frac{1}{N} \sum_{j=1}^{N} \frac{1}{L - \lceil L/2 \rceil} \min \left\{ l(z_j), L - l(z_j) \right\}, \tag{6}$$

where L is the number of classifiers, N is the number of instances in the data set. z_j is instance. $l(z_j)$ is the number of classifiers that can correctly recognize z_j at the same time. E varies between 0 and 1. Except these above, we use pairs of diversity measures, as seen [5].

3. Algorithms of selective ensemble learning

Selective Ensemble learning is to select ensemble members with selective strategies after generating many different base models. Different selective strategies can get different ensemble learning algorithms. In the past, researchers mainly improve diversity of ensemble member by use feature subset method [6-11]. In this paper, we use data subset method and give four different selective ensemble learning approaches: Hill-Climbing, Ensemble Backward Sequential Selection, Ensemble Forward Sequential Selection and Clustering Selection. Base models are support vector machine (SVM).

During training classifiers, classifiers have differences because they are trained by randomly extracting data set. In every training classifier, the solution space is different. In order to measure how the differences impact on the accuracy of classifier, we introduce a formula to study how diversity impact on ensemble accuracy. It defines as follows:

$$Fun = \frac{acc}{allacc} + r \cdot \frac{div}{alldiv}. \tag{7}$$

In formula (7), *acc* is ensemble accuracy. *allacc* is all classifiers' accuracy. *div* is ensemble diversity. *alldiv* is all classifiers' diversity. In computing accuracy, we use the majority vote and use the entropy E as the diversity measure. In the following, we breifly introduce four different ensemble learning methods which are seen as in [4].

3.1 Hill-Climbing (HC) Method

Hill Climbing ensemble proposed is composed of two major phases, namely construction of the initial ensemble by randomly selecting base model and iterative refinement of the ensemble members. Initial ensemble members are formed using the randomly selective method; the second phase is aimed to improve the value of the fitness function of the ensemble classifiers. For all the learning models, an attempt is made to switch (add or delete) each models. If the result produces the larger value of fitness, that change is kept. This process is continued until no further improvements are possible.

3.2 Ensemble Backward Sequential Selection (EBSS) Method

EBSS begins with all learning models and repeatedly removes a model whose removal yields the maximal value of fitness improvement. The cycle repeats until no improvement is obtained.

3.3 Ensemble Forward Sequential Selection (EFSS) Method

EFSS begins with zero attributes, evaluates all base models with exactly one model, and selects the one with the best performance. It then adds to the ResultSet that yields the best performance for models of the next larger size. The cycle repeats until no improvement is obtained.

3.4 Clustering Selection Ensemble

Clustering technology is an important data analysis tool. By it, data structure may be found. At present, there exist many different kinds of clustering algorithms. Among them, most common clustering algorithms are hierarchical clustering Algorithms and k-means clustering algorithms. In the following, we study model clustering based on above algorithms.

For any two models n_m and n_n, distance between them is defined as

$$\forall n_m, n_n \in E \quad d(n_m, n_n) = \frac{1}{N}\sum_{j=1}^{N} \frac{1}{L - \lceil L/2 \rceil} \min\left\{ l(z_j), L - l(z_j) \right\}. \tag{8}$$

The above distance measure is aimed to group models based on diversity. That is to say that in the same cluster, we select base model so that the value of diversity in the whole cluster is maximal. Moreover, in hierarchical clustering algorithms, to merge similar clusters, we use the following distance between any two clusters:

$$\forall E_i, E_j \quad i \neq j \quad d(E_i, E_j) = \max_{n_s \in E_i, n_t \in E_j} \{ d(n_s, n_t) \}. \tag{9}$$

In the following, T is the size of training models, S is dataset and L is learning algorithm. The selective ensemble method of hierarchical clustering is described as follows.

4. Experiments

4.1 Experimental data and methods

A number of ensemble techniques solving the integration problem can improve the generalization performance of ensemble learning. In addition, the theoretical basis of ensemble learning is based on the diversity of base models. Selective integration aims to select ensemble models which have the biggest diversity using some strategies. In this paper these strategies include Hill Climbing, Ensemble Backward Sequential Selection, Ensemble Backward Sequential Selection and Clustering technology. Clustering technology include hierarchical clustering algorithms and k-means clustering. These methods have in common, that is, many base models are all trained utilizing decision tree algorithms, neural network algorithm and support vector algorithm before getting ensemble models. Then, ensemble models are constructed with above selective ensemble methods.

TABLE I Features of dataset

Number	Data set	Number of data	Number of feature	Number of class
1	Balance	625	5	3
2	Car	1728	6	4

3	Cmc	1473	9	3
4	Ecoli	336	8	8
5	Glass	214	11	7
6	Hayes	132	6	3
7	Iris	150	4	3
8	Pima	768	8	2
9	Wine	178	13	3
10	Zoo	101	18	7

4.2 .Experimental results and analysis for HC, EFSS and EBSS

Firstly, many base models are trained with decision tree, BP neural network and support vector machine algorithm. Then we use hill climbing, ensemble forward sequential selection and ensemble backward sequential selection to select base models for forming ensemble members. Finally, the performance of integrated model and diversity are achieved using a majority vote.

In order to study the effect of the diversity on ensemble accuracy, we consider the accuracy and diversity together. It controls diversity by altering parameter *r*. The value of parameter *r* is 0, 1/5, 1/3 and 1. In Fig. 1 and Fig. 2, classifiers are trained using support vector machine. Then ensemble members are achieved using above three selective methods. Finally ensemble accuracy is computed using vote method. From Fig. 2, we can see that when parameter r is zero, the value of most ensemble accuracy is bigger than other. For some data set, the ensemble accuracy is larger as the value of parameter *r*. So the diversity impact on the ensemble accuracy in some extent. In all, three different ensemble methods including HC,EFSS and EBSS have similar generalization performance based on support vector machine for vote strategy.

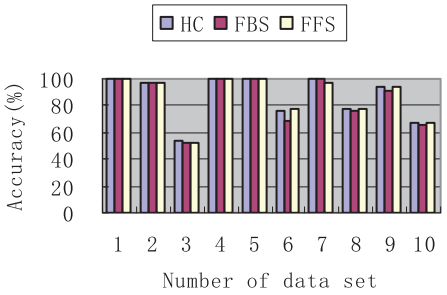

Fig. 1 Expermental results with HC,EFSS and EBSS methods.

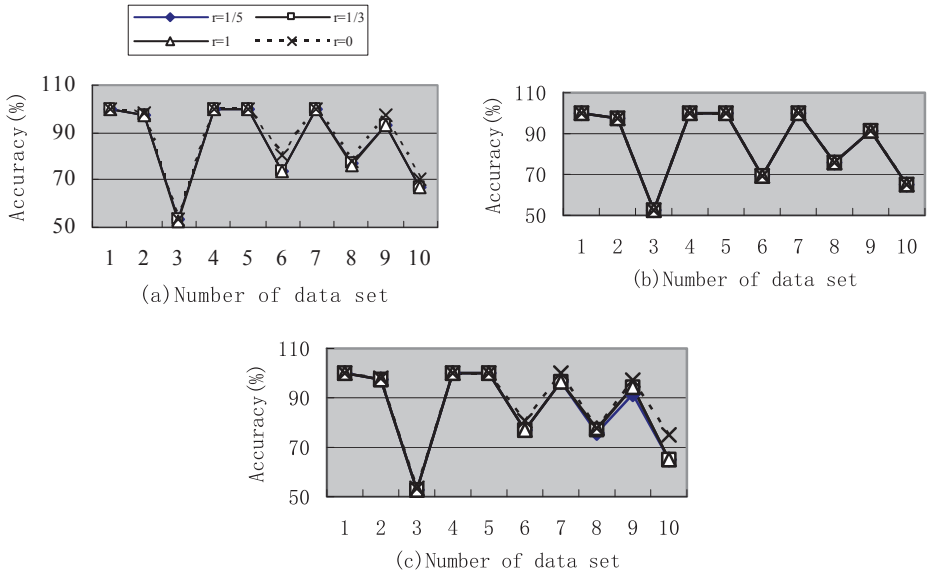

Fig. 2 Expermental results with HC,EFSS and EBSS methods for different value r.

4.3 Experimental results with clustering technology

A hundred of base models are created using the support vector machine algorithm. Then different clusters are obtained using k-means and hierarchical clustering technology. In this experiment, the number of clusters is 4, 6, 9, 15, 25, 30 and 35. Fig. 3 is the average result of ensemble models which is selected using k-means clustering technology and hierarchical clustering technology for different cluster. Ensemble accuracy is computed by vote method, and diversity is measured by entropy E, Fail/Non-fail, double fault and plain disagreement measure. In clustering selective technology, we select classifiers which have larger diversity. From Fig. 3, we can see that there is no clear disciplinary change, but using clustering strategy can get better ensemble performance.

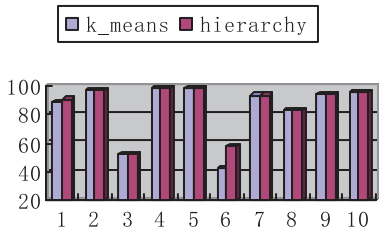

Fig. 3 Expermental results with Clustering methods.

4.4 Analysis of experimental results with generalization error

We consider the generalization error of ensemble learning when generating the outputs of a number of different classifiers. With regard to the generalization error, Martin studied it [14]. In the experiment, we first select some base models using clustering technology including k-means and hierarchical clustering technology, then using these base models estimate the generalization error of test data. We regard the data set as a sample $s = ((x_1, b_1), ..., (x_m, b_m)) \in Z^m$ and $b \in \{0,1\}$. Suppose that selected base models are $\{C_1, C_2, \ldots, C_L\}$. With a test sample x_j, we compute $Count_j = |\{C_i : C_i(x_j) = b_j\}| \quad i = 1, 2, ..., L$, thus there is following equation:

$$er_j = \begin{cases} 1 & Count_j > L/2 \\ 0 & Count_j < L/2 \end{cases}.$$

Finally, all er_j are sumed as follows:

$$er_{count} = \sum_{i=1}^{L} er_i .$$

The generalization error is,

$$error = \frac{er_{count}}{m} .$$

Experimental result is seen as in Fig. 4. From these result, we can see that selective strategies can reduce ensemble generalization error. Further, they can improve ensemble generalization performance.

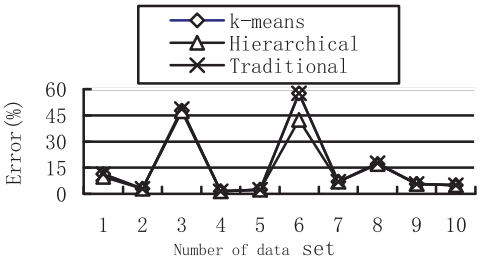

Fig. 4 Expermental results with Clustering methods.

5. Conclusions

The paper primarily studied the selective ensemble methods and the diversity of ensemble models. We first introduce diversity measures including pairs of diversity measures and non-pairwise diversity measures. Pairs of diversity measures include fail/non-fail, double-fault and plain disagreement measure. Non-pairwise diversity measure is the entropy measure E. Then, we study four selective ensemble technologies, namely hill climbing, ensemble forward sequential selection, ensemble backward sequential selection and clustering technologies including hierarchical clustering algorithms and k-means clustering. The Entropy E is used as diversity measures in this paper. Analyzing the hill climbing, ensemble forward sequential selection and ensemble backward sequential selection, we can see that using selecting strategy can achieve a certain performance advantages of integration.

6. Acknowledgment

The authors would like to thank Hebei Natural Science Foundation for its financial support (No: F2009000236).

References

[1] Giacinto, G., Roli, F. Design of effective neural network ensembles for image classification processes. Image Vision and Computing Journal, 2001, 19(9/10): 699~707.

[2] Zhi Hua Zhou, Jianxin Wu, Wei Tang. Ensembling neural networks : Many could be better than all .Artificial Intelligence, 2002 , 137 (1/2) : 239~263.

[3] Li Guo-zheng, Yang Jie etal.Clustering algorithms based selective ensemble. Journal of Fudan university(Natural Science),2004,43(5):689~691.

[4] Li Kai, Han Yanxia. Study of selective ensemble learning method and its diversity based on decision tree and neural network. 2010 Chinese Control and Decision Conference, CCDC 2010,p1310-1315.

[5] L. I. Kuncheva and C. J. Whitaker. Measures of diversity in classifier ensembles. Machine Learning, 2003,51:181~207.

[6] Alexey Tsymbal. Mykola Pechenizkiy. Padraig Cunningham. Diversity in search strategies for ensemble feature selection. Information Fusion 6(2005) 83~98.

[7] A.Tsymbal, S.Puuronen, D. Patterson. Ensemble feature selection with the simple Bayesian classification.Information Fusion,2003,4(2) :87~100.

[8] P. Delimata, Z. Suraj, Feature Selection Algorithm for Multiple Classifier Systems:A Hybrid Approach. Fundamenta Informaticae, 2008,85:97~110.

[9] Eulanda M. Dos Santos, Robert Sabourin, Patrick Maupin. A dynamic overproduce-and-choose strategy for the selection of classifier ensembles. Pattern Recognition,2008,41:2993~3009.

[10] Eulanda M. Dos Santos, Robert Sabourin, Patrick Maupin. Overfitting cautious selection of classifier ensembles with genetic algorithms. Information Fusion, 2009,10:150~162.

[11] G. Martinez-Munoz, A. Suarez, Using boosting to prune bagging ensembles.Pattern Recognition Letters,2007,28 (1) :156~165.

[12] Ioannis Partalas, Grigorios Tsoumakas and Ioannis Vlahavas. Pruning an Ensemble of Classifiers via Reinforcement Learning. Neurocomputing, 2009,72 (7-9):1900~1909.

[13] D. Skalak. The sources of increased accuracy for two proposed boosting algorithms. American Association for Artificial Intelligence. AAAI-96 (1996).

[14] A. Martin. On the generalization error of fixed combinations of classifiers. Journal of Computer and System Sciences 73 (2007) 725-734.